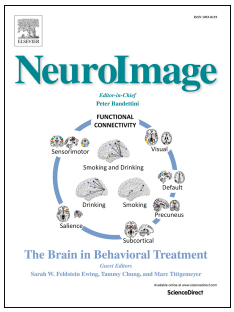
Accepted Manuscript



What do results from coordinate-based meta-analyses tell us?

Anton Albajes-Eizagirre, Joaquim Radua

PII: S1053-8119(18)30385-9

DOI: 10.1016/j.neuroimage.2018.04.065

Reference: YNIMG 14912

To appear in: *NeuroImage*

Received Date: 15 January 2018

Revised Date: 27 April 2018

Accepted Date: 28 April 2018

Please cite this article as: Albajes-Eizagirre, A., Radua, J., What do results from coordinate-based meta-analyses tell us?, *NeuroImage* (2018), doi: 10.1016/j.neuroimage.2018.04.065.

What do results from coordinate-based meta-analyses tell us?

Anton Albajes-Eizagirre[1,2] and Joaquim Radua[1-4]

[1] FIDMAG Germanes Hospitalàries, Sant Boi de Llobregat, Barcelona, Spain

[2] Mental Health Research Networking Center (CIBERSAM), Madrid, Spain

[3] Centre for Psychiatric Research and Education, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

[4] Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

Running title: Results from coordinate-based meta-analyses

Correspondence to:
Joaquim Radua
King's College London, Institute of Psychiatry, Psychology and Neuroscience
PO 69. Division of Psychosis Studies
16 De Crespigny Park, London, SE5 8AF
Telephone: 02078480363 - FAX: 02078480379
Email: *quimradua@gmail.com*

Keywords: coordinate-based meta-analysis, tests for spatial convergence, familywise error rate, activation likelihood estimation, seed-based d mapping, signed differential mapping

Number of words in the abstract: 190
Number of words in the text: 1984
Number of figures: 1
Number of tables: 0

ABSTRACT

Coordinate-based meta-analyses (CBMA) methods, such as Activation Likelihood Estimation (ALE) and Seed-based d Mapping (SDM), have become an invaluable tool for summarizing the findings of voxel-based neuroimaging studies. However, the progressive sophistication of these methods may have concealed two particularities of their statistical tests. Common univariate voxelwise tests (such as the t/z-tests used in SPM and FSL) detect voxels that activate, or voxels that show differences between groups. Conversely, the tests conducted in CBMA test for "spatial convergence" of findings, i.e., they detect regions where studies report "more peaks than in most regions", regions that activate "more than most regions do", or regions that show "larger differences between groups than most regions do". The first particularity is that these tests rely on two spatial assumptions (voxels are independent and have the same probability to have a "false" peak), whose violation may make their results conservative or liberal, though fortunately current versions of ALE, SDM and some other methods consider these assumptions. The second particularity is that the use of these tests involves an important paradox: the statistical power to detect a given effect is higher if there are no other effects in the brain, whereas lower in presence of multiple effects.

1. Introduction

The exponential increase of voxel-based neuroimaging studies led to the need of methods that could summarize their results. Neuroimaging papers usually only report coordinates and statistics of the peaks (or "foci") of the clusters of statistical significant voxels, and thus data extracted from these studies are a series of numeric tables that classical meta-analytic methods cannot combine. In this context, several developers introduced methods for conducting coordinate-based meta-analyses (CBMA), which are able to integrate this wealth of numeric information and return clear summary brain maps, thus shedding light on the neural substrates of many brain functions and neuropsychiatric disorders. Examples of these methods are Activation Likelihood Estimation (ALE) and Seed-based d Mapping (SDM), among others [1-16].

Importantly, the statistical tests used by these methods have two particularities as compared to the common univariate voxelwise tests present in neuroimaging software such as SPM or FSL. Univariate voxelwise tests, which may be used at subject-level, group-level or even study-level (e.g., to conduct a classic meta-analysis when all study data are available), assess whether a voxel shows not-null activation (i.e. blood oxygenation level dependent –BOLD– response is not zero), or whether a voxel shows not-null differences between groups (i.e. values in the two groups are not identical). Their statistics (usually t/z-values) summarize evidence against the null hypotheses "absence of BOLD response or differences between groups". Conversely, the tests conducted in CBMA test for "spatial convergence" of findings, i.e. they assess whether studies report more findings in the neighborhood of a given voxel than in the neighborhood of most voxels [8]. We show here that these tests rely on two spatial assumptions, whose violation may make their results conservative or liberal, and that their statistical power decreases when there are multiple effects in the brain. We first present a toy meta-analysis to help us illustrate these points.

2. A toy meta-analysis

For simplicity, we may imagine that the gray matter mask is composed of several independent voxels. The values of these voxels may be random t-values converted into effect sizes [6]. Voxels whose values reach a given threshold may be considered "peaks" and set to "one", whereas the value of the remaining voxels may be set to "zero". The toy meta-analysis may simply consist of calculating the mean of the studies, separately for each voxel.

A test for spatial convergence could consist of repeatedly permuting the values "between the voxels" (i.e. randomizing the location of the peaks), to simulate meta-analyses in which any spatial convergence is only due to chance. The means of these permuted data would compose a null distribution from which we can derive the probability to obtain means as high as the original means by chance (i.e. the p-values). Specifically, each permutation would include the following steps: a) randomly swapping the effect-sizes between the voxels separately for each study; b) recalculating the means of the permuted studies, separately for each voxel; and c) saving the maximum of the means. If we aimed to control the familywise error rate (FWER) at 5% (i.e., to have a probability of 5% of making one or more type I errors), we would consider a voxel statistically significant if its original mean was higher than 95% of these maxima. The reader may see Figure 1A and run Simulation 1 (Supplement) for an example. Note that for simplicity, the figure includes only six voxels, but the reader may set as many hundreds or thousands of voxels as desired in the simulations.

The reason why this test only saves the maxima is not related to CBMA but to the correction for multiple comparisons [17]. It is obvious that 5% of the meta-analyses simulated in the permutations would have maxima that are higher than 95% of the maxima, and as we would wrongly consider these meta-analyses statistically significant, the FWER would be 5% (as we wish). The choice of FWER = 5% is arbitrary, other significance levels may be used.

Conversely, we would ask the reader to focus on how the test conducts a permutation: the null hypothesis is that peaks are randomly located within the gray matter mask, and to this end, we randomly reallocate the peaks during the permutations.

This procedure is radically different from the voxelwise permutations tests, such as those used in FSL "randomize", which do not swap voxel values [18]. In a one-sample permutation, these tests multiply a random set of the individual images by -1, and in a two-sample permutation, they randomly reassign the individuals to the two samples [18]. Afterwards, they recalculate the test statistics (e.g. t/z-values). For instance, to infer whether there are brain activation differences between males and females, these tests would first calculate the t-value image of the comparison between our sample of males (e.g., David, John and Robert) and our sample of females (e.g., Tina, Mary and Linda). In the first permutation, the tests could randomly assign Mary, David and Linda to the "male" group, and John, Tina and Robert to the "female group", and they would re-calculate the t-value image. In the second permutation, they could randomly reassign Linda, John and David to the "male" group, and Robert, Tina and Mary to the "female group", and they would re-calculate the t-value image again. And so on. With these random reassignments, the permutation tests break any association between brain activation and group labels.

3. Spatial assumptions

In the unrealistic toy meta-analysis, the permutation test was accurate because i) each voxel was independent from its neighbors, and ii) each voxel had the same probability to have a "false" peak. However, the data may not always meet these assumptions, as detailed in the following.

First, in real gray matter, voxels correlate with their neighbors, local peaks from the same cluster tend to be very close, peaks from neighboring related clusters are closer than peaks from independent clusters, and etcetera. If the data simulated in the permutations do not have the spatial structure of the original data, there are differences between the original data and the permuted data that are unrelated to spatial convergence but due to the differences in spatial structures. For instance, two local peaks from the same cluster are usually very close, whereas in the permutations they could be at any distance. The reader may run Simulation 2 (Supplement) for an example where the data do not meet the assumption of spatial independence because the peaks are very close, simulating close local peaks from the same cluster. In this example, the test would be substantially conservative.

Moreover, the destruction of the spatial structure in the permutations invalidates the use of spatial statistics, in which p-values are derived from the cluster sizes (or similar measures such as cluster masses or TFCEs [19]). The spatial correlation between neighbor voxels involves that statistically significant voxels tend to be together forming clusters. If these correlations are not present in the permutations, statistically significant voxels are sparse and do not form clusters during the permutations, inflating the statistical significance (the p-values associated to the different cluster sizes become too small, i.e., clusters as large as the ones observed in the unpermuted data would be extremely unlikely in the permuted data).

To preserve the spatial structure, permutation tests should ensure that effect sizes from neighbor voxels remain together, or that the Euclidean distances between peaks are unmodified, in all permutations. Multilevel Kernel Density Analysis (MKDA) introduced the swapping of blocks of voxels, rather than voxels, in an attempt to preserve this spatial structure [10], and similar approaches were subsequently added to ALE [4] and SDM [6].

Second, probably all studies cover voxels that are mostly composed of gray matter, but only some of the studies may cover voxels that are only partially composed of gray matter. If only some of the studies cover a voxel, it is obviously less likely that a study reports a peak in this voxel, violating the assumption

of homogeneity of the probability to have a "false" peak. The reader may run Simulation 3 (Supplement) for an example where this violation would make the test substantially liberal. Fortunately, some modern CBMA methods such as SDM include accurate tissue templates to minimize this effect [20, 21].

These issues do not apply to the voxelwise permutations tests [18] because they do not swap values between voxels.

4. Statistical power and number of effects

Imagine that all studies in the toy meta-analysis reported a peak in the first voxel and no other findings. In the permutations, each study would have one peak randomly located in any of the voxels, and thus the probability that a voxel of a study had a peak would be $1/N_{voxels}$. The probability that this voxel had a peak in all studies would be $1/N_{voxels}$ raised to $N_{studies}$. Finally, we could multiply this probability and the number of voxels to have the probability that any voxel had a random peak in all studies, i.e. the probability to have a meta-analytic value as high as the meta-analytic value in the original data:

$$P_1 = \left(\frac{1}{N_{voxels}}\right)^{N_{studies}} \cdot N_{voxels} = \left(\frac{1}{N_{voxels}}\right)^{N_{studies}-1}$$

Now, imagine that in addition to the peak in the first voxel, all studies reported a peak in the second voxel. Following a similar argument, we could derive that the probability to obtain, in a permutation, a meta-analytic value as high as the meta-analytic values in the original data would be:

$$P_2 = \left(\frac{2}{N_{voxels}}\right)^{N_{studies}} \cdot N_{voxels} - \left(\frac{1}{\binom{N_{voxels}}{2}}\right)^{N_{studies}} \cdot \binom{N_{voxels}}{2} = \left[2^{N_{studies}-1} \cdot \left(2 - \frac{1}{(N_{voxels}-1)^{N_{studies}-1}}\right)\right] \cdot P_1$$

Given that the expression between square brackets is substantially larger than one, P_2 is substantially larger than P_1, i.e. there would be a substantial increase in the probability that a voxel has a meta-analytic value as high as the meta-analytic values in the original data, and this increase involves a substantial decrease in statistical power. The user may run Simulations 3 and 4 (Supplement) for an example of substantially lower statistical power when the same test is conducted in the presence of a single effect than when is conducted in the presence of multiple effects. Interestingly, these simulations show that the reduction of power might be stronger when the number of voxels is larger.

Figure 1B also shows how the threshold substantially increases (reducing statistical power) in the presence of multiple effects. In the left example, 50% of the studies have one true peak and no false peaks, and this true peak is statistically significant (its value, 0.5, is larger than the threshold, 0.3). In the right example, 50% of the studies have three true peaks and no false peaks, and these true peaks are not statistically significant (their value, 0.5, is lower than the threshold, 0.6).

The situation is different in univariate voxelwise tests, because the uncorrected p-value of one voxel does not depend on the values of the other voxels (beyond the correlation expected between correlated voxels, e.g. in real data adjacent voxels have similar p-values).

5. Are the tests comparable?

Whether these tests are comparable and to what extent, may be a matter of discussion. On the one hand, they have different uses, as group analyses of individual images use common voxelwise tests, while CBMA use tests for spatial convergence. However, this association might simply be due to the limited availability of data for voxel-based meta-analyses, because voxel-based meta-analyses can use common voxelwise tests if all study data are available. And vice versa: there is indeed no theoretical impediment to use a test for convergence to conduct a group-level test (i.e., looking for the convergence of the peaks found in the subject-level tests). On the other hand, common voxelwise tests have a set of steps that result in a threshold that ensures that, in the absence of true effects, there is only 5% probability that a voxel is statistically significant, and tests for spatial convergence have different steps and result in a different threshold that ensures that, if activations or differences are distributed uniformly throughout the gray matter, there is only 5% probability that a voxel is statistically significant. In other words, researchers can conclude that voxels with values above the voxelwise-test-threshold "activate" (or show differences between groups), whereas voxels with values above the test-for-convergence-threshold "activate more than most voxels do" (or show larger differences between groups than most voxels do). Therefore, the tests may have a common goal, but are used in different scenarios and assess indeed different things.

Acknowledgements

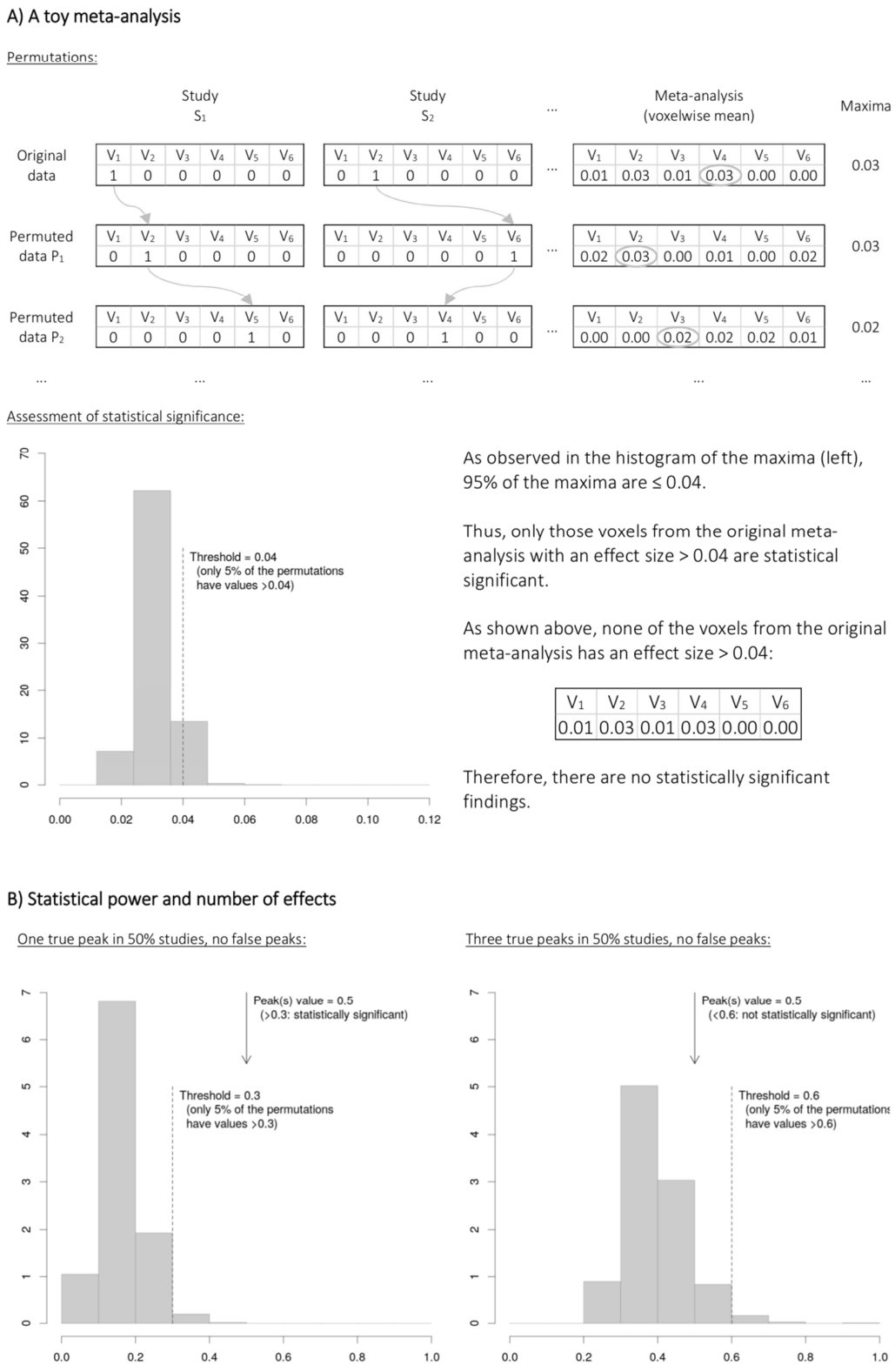
This work was supported by Miguel Servet Research Contract MS14/00041 and Research Project PI14/00292 from the Plan Nacional de I+D+i 2013–2016, the Instituto de Salud Carlos III-Subdirección General de Evaluación y Fomento de la Investigación and the European Regional Development Fund

(FEDER). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

References

1. Turkeltaub, P.E., et al., *Meta-analysis of the functional neuroanatomy of single-word reading: method and validation.* Neuroimage, 2002. **16**(3 Pt 1): p. 765-80.
2. Eickhoff, S.B., et al., *Activation likelihood estimation meta-analysis revisited.* Neuroimage, 2012. **59**(3): p. 2349-61.
3. Laird, A.R., et al., *ALE meta-analysis: controlling the false discovery rate and performing statistical contrasts.* Hum Brain Mapp, 2005. **25**(1): p. 155-64.
4. Eickhoff, S.B., et al., *Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty.* Hum Brain Mapp, 2009. **30**(9): p. 2907-26.
5. Turkeltaub, P.E., et al., *Minimizing within-experiment and within-group effects in Activation Likelihood Estimation meta-analyses.* Hum Brain Mapp, 2012. **33**(1): p. 1-13.
6. Radua, J., et al., *A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps.* Eur Psychiatry, 2012. **27**(8): p. 605-11.
7. Radua, J. and D. Mataix-Cols, *Meta-analytic methods for neuroimaging data explained.* Biol Mood Anxiety Disord, 2012. **2**: p. 6.
8. Radua, J. and D. Mataix-Cols, *Voxel-wise meta-analysis of grey matter changes in obsessive-compulsive disorder.* Br J Psychiatry, 2009. **195**(5): p. 393-402.
9. Radua, J., et al., *Anisotropic kernels for coordinate-based meta-analyses of neuroimaging studies.* Front Psychiatry, 2014. **5**: p. 13.
10. Wager, T.D., M. Lindquist, and L. Kaplan, *Meta-analysis of functional neuroimaging data: current and future directions.* Soc Cogn Affect Neurosci, 2007. **2**(2): p. 150-8.
11. Costafreda, S.G., A.S. David, and M.J. Brammer, *A parametric approach to voxel-based meta-analysis.* Neuroimage, 2009. **46**(1): p. 115-22.
12. Kang, J., et al., *Meta Analysis of Functional Neuroimaging Data via Bayesian Spatial Point Processes.* J Am Stat Assoc, 2011. **106**(493): p. 124-134.
13. Yue, Y.R., M.A. Lindquist, and J.M. Loh, *Meta-analysis of functional neuroimaging data using Bayesian nonparametric binary regression.* Ann. Appl. Stat., 2012. **6**(2): p. 697-718.
14. Kang, J., et al., *A Bayesian Hierarchical Spatial Point Process Model for Multi-Type Neuroimaging Meta-Analysis.* Ann Appl Stat, 2014. **8**(3): p. 1800-1824.
15. Montagna, S., et al., *Spatial Bayesian latent factor regression modeling of coordinate-based meta-analysis data.* Biometrics, 2017.
16. Tench, C.R., et al., *Coordinate based random effect size meta-analysis of neuroimaging studies.* Neuroimage, 2017. **153**: p. 293-306.
17. Nichols, T.E. and A.P. Holmes, *Nonparametric permutation tests for functional neuroimaging: a primer with examples.* Hum Brain Mapp, 2002. **15**(1): p. 1-25.
18. Winkler, A.M., et al., *Permutation inference for the general linear model.* Neuroimage, 2014. **92**: p. 381-97.
19. Smith, S.M. and T.E. Nichols, *Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference.* Neuroimage, 2009. **44**(1): p. 83-98.
20. Radua, J., et al., *Voxel-based meta-analysis of regional white-matter volume differences in autism spectrum disorder versus healthy controls.* Psychol Med, 2011. **41**(7): p. 1539-50.
21. Peters, B.D., et al., *White matter development in adolescence: diffusion tensor imaging and meta-analytic results.* Schizophr Bull, 2012. **38**(6): p. 1308-17.

Figure 1: A toy meta-analysis



Conflict of interest statement

This work was supported by Miguel Servet Research Contract MS14/00041 from the Plan Nacional de I+D+i 2013–2016, the Instituto de Salud Carlos III-Subdirección General de Evaluación y Fomento de la Investigación and the European Regional Development Fund (FEDER). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

The author reports no conflicts of interests related to this manuscript.